Sequential Coding of Gauss–Markov Sources

Anatoly Khina, Ashish Khisti, Victoria Kostina, and Babak Hassibi

***Abstract*—We consider the problem of sequential transmission of Gauss–Markov sources. We show that in the limit of large spatial block lengths, greedy compression with respect to the squared error distortion is optimal; that is, there is no tension between optimizing the distortion of the source in the current time instant and that of future times. We then extend this result to the case where at time t a random compression rate R_t is allocated independently of the rate at other time instants. This, in turn, allows us to derive the optimal performance of sequential coding over packet-erasure channels with instantaneous feedback. For the case of packet erasures with delayed feedback, we connect the problem to that of compression with side information that is known at the encoder and may be known at the decoder — where the most recent packets serve as side information that may have been erased. We conclude the paper by demonstrating that the loss due to a delay by one time unit is rather small.**

***Index Terms*—Sequential coding of correlated sources, successive refinement, source streaming, packet erasures, source coding with side information.**

I. Introduction

Sequential coding of sources is increasingly finding applications, such as real-time video streaming, and cyberphysical and networked control. Such systems use compressed packet-based transmission and strive to achieve minimum distortion for the given compression rates.

This setting was introduced and treated for the two-source case by Viswanathan and Berger [1] and for more users in [2]–[5]. For the special case of Gauss–Markov sources, an explicit expression for the achievable sum-rate for given distortions was derived in [2], [3] and extended for the (general) jointly Gaussian three-source case in [6].

In practice, however, packet-based protocols are prone to erasures and possible delays. The case of sequential coding in the presence of packet erasures was treated for various erasure models. The case when only the first packet is prone to an erasure was considered in [7]. A more general approach which trades between the performance given all previously sent packets and the performance given only the last packet was proposed in [8]. For random independent identically distributed (i.i.d.) packet erasures, a hybridation between pulse-code modulation (PCM) and differential PCM (DPCM), termed leaky DPCM, was proposed in [9] and analyzed for the case of very low erasure probability in [10]. The scenario in which the erasures occur in bursts was considered in [11], [12]. There, a sequence of source vectors sampled from a Gauss–Markov process in the temporal dimension must be encoded sequentially and reconstructed with zero delay at the decoder. The channel introduces a burst of erasures of a certain maximum length and the decoder is not required to reconstruct the sequences that fall in the erasure period and a recovery window following it.

All of these works assume no feedback is available at the encoder, namely that the encoder does not know whether a transmitted packet successfully arrives to the decoder or is erased in the process.

In this paper, we first consider the problem of sequential coding of Gauss–Markov sources and determine the rate–distortion region for large frames. Specifically, we show that greedy quantization that optimizes the distortion for each time is also optimal for minimizing the distortion of future time instants. This insight allows us to extend the result to the case where the compression rate r_t available for the transmission of the packet at time t is determined just prior to its transmission.

The packet-erasure channel with instantaneous output feedback (ACK/NACK) can be viewed as a special case of the above noiseless channel with random rate allocation, with $r_t = 0$ corresponding to a packet-erasure event [13]. The optimal rate–distortion region of sequential coding of Gauss–Markov sources in the presence of packet erasures and instantaneous output feedback thereby follows as a simple particularization of our more general result.

We further tackle the more challenging delayed feedback setting, in which the encoder does not know whether the most recently transmitted packets arrived or not. Viewing these recent packets as side information (SI) that is available at the encoder and possibly at the decoder, and leveraging the results of Kaspi [14] along with their specialization for the Gaussian case by Perron *et al.* [15],[1] we adapt our transmission scheme to the case of delayed feedback. We provide a detailed description of the proposed scheme for the case where the feedback is delayed by one time unit and demonstrate that the loss compared to the case of instantaneous feedback is small.

II. Problem Statement

We now present the model of the source, channel, and the admissible encoder and decoder both of which are required to be causal in this work; see Fig. 1.

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm. Random variables are denoted by lower-case letters with

A. Khina, V. Kostina and B. Hassibi are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA. E-mails: {*khina, vkostina, hassibi*}*@caltech.edu*

A. Khisti is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada. E-mail: *akhisti@comm.utoronto.ca*

[1]The scenario considered in [14], [15] can be also viewed as special case of the results of Heegard and Berger [16], where the SI is not available at the encoder, by adjusting the distortion measure and "augmenting" the source [17]. Interestingly, knowing the SI at the encoder allows one to improve the optimal performance of this scenario in the Gaussian case; see Rem. 9.

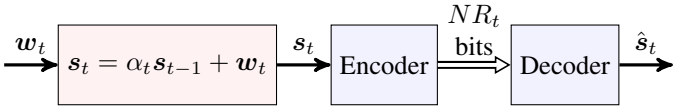


Fig. 1: Sequential coding of a Gauss–Markov source setup.

temporal subscripts (a_t), and random vectors ("frames") of length N by boldface possibly accented lower-case letters with temporal subscripts ($\boldsymbol{a}_t, \hat{\boldsymbol{a}}_t$). We denote temporal sequences by $\boldsymbol{a}^t \triangleq (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_t)$. $\mathbb{N}$ is the set of natural numbers. All other notations represent deterministic scalars.

We assume that the communication spans the time interval $[1, T]$, where $T \in \mathbb{N}$.

Source: Consider a Gauss–Markov source $\{\boldsymbol{s}_t\}$, whose outcomes are vectors ("*frames*") of length N with i.i.d. samples along the spatial dimension, that satisfy the temporal Markov relation:

$$\boldsymbol{s}_t = \alpha_t \boldsymbol{s}_{t-1} + \boldsymbol{w}_t, \qquad t = 1, \ldots, T\,, \tag{1}$$

where $\{\alpha_t\}$ are known process coefficients that satisfy $|\alpha_t| < 1$, and the outcomes of $\{\boldsymbol{w}_t\}$ are i.i.d. along the spatial dimension, Gaussian and mutually independent across time of zero mean and variances $\{W_t\}$. We assume $\boldsymbol{s}_0 = 0$ for convenience.

Denote by $S_t \triangleq \frac{1}{N}\mathbb{E}\left[\|\boldsymbol{s}_t\|^2\right]$ the average power of the entries of vector $\boldsymbol{s}_t$. Then, we obtain the following recursive relation:

$$S_t = \alpha_t^2 S_{t-1} + W_t, \qquad t = 1, \ldots, T\,, \tag{2a}$$
$$S_0 = 0. \tag{2b}$$

Channel: At time t, a packet $f_t \in \left\{1, 2, \ldots, 2^{NR_t}\right\}$ is sent over a noiseless channel of finite rate R_t.

Causal encoder: Sees $\boldsymbol{s}_t$ at time t and applies a causal function $\mathscr{F}_t$ to the entire observed source sequence $\boldsymbol{s}^t$ to generate a packet $f_t \in \left\{1, 2, \ldots, 2^{NR_t}\right\}$:

$$f_t = \mathscr{F}_t\left(\boldsymbol{s}^t\right). \tag{3}$$

Causal decoder: Applies a causal function $\mathscr{G}_t$ to the sequence of received packets f^t to construct an estimate $\hat{\boldsymbol{s}}_t$ of $\boldsymbol{s}_t$, at time t:

$$\hat{\boldsymbol{s}}_t = \mathscr{G}_t\left(f^t\right). \tag{4}$$

Distortion: The mean-square error distortion at time t is defined as

$$D_t \triangleq \frac{1}{N}\mathbb{E}\left[\|\boldsymbol{s}_t - \hat{\boldsymbol{s}}_t\|^2\right]. \tag{5}$$

If we specialize the source process into that of fixed parameters, namely,

$$\begin{aligned} \alpha_t &\equiv \alpha, \\ W_t &\equiv W, \end{aligned} \qquad t = 1, \ldots, T\,, \tag{6}$$

then its power converges to

$$S_\infty = \frac{W}{1 - \alpha^2}\,.$$

We further define the steady-state distortion (assuming the limit exists):

$$D_\infty \triangleq \lim_{T \to \infty} D_t\,.$$

Definition (Distortion-rate region)**.** The *distortion–rate region* is the closure of all achievable distortion tuples $D^T \triangleq (D_1, \ldots, D_T)$ for a rate tuple $R^T \triangleq (R_1, \ldots, R_T)$, for any N, however large; its inverse is the *rate–distortion region*.

III. Distortion–Rate Region of Sequential Coding of Gauss–Markov Sources

The optimal achievable distortions for given rates for the model of Sec. II are provided in the following theorem.

Theorem 1 (Distortion–rate region)**.** *The distortion–rate region of sequential coding for a rate tuple R^T is given by all distortion tuples D^T that satisfy $D_t \geq D_t^*$ with*

$$D_t^* = \left(\alpha_t^2 D_{t-1}^* + W_t\right) 2^{-2R_t}, \qquad t = 1, \ldots, T\,, \tag{7a}$$
$$D_0^* = 0. \tag{7b}$$

Remark 1. Th. 1 establishes the optimal rate–distortion region for the "causal encoder–causal decoder" setting of Ma and Ishwar [2] for the case of Gauss–Markov sources. We note that Ma and Ishwar [2] provide an explicit result only for the sum-rate for the Gauss–Markov case [3]. Torbatian and Yang [6] extend the sum-rate result to the case of three jointly Gaussian sources (which do not necessarily constitute a Markov chain). Our work, on the other hand, fully characterizes the rate–distortion region for the case of Gauss–Markov sources.

Remark 2. The results and proof (provided in the sequel) of Th. 1 imply that optimal greedy quantization at every step — which is achieved via Gaussian backward [18, Ch. 10.3] or forward [18, pp. 338–339] channels — becomes optimal when N is large. Moreover, it achieves the optimum for all $t \in [1, T]$ simultaneously, meaning that there is no tension between minimizing the current distortion and future distortions.

To prove this theorem we first construct the optimal greedy scheme and determine its performance in Sec. III-A. We then show that it is in fact (globally) optimal when N goes to infinity, by constructing an outer bound for this scenario, in Sec. III-B.

A. Achievable

We construct an inner bound using the optimal greedy scheme. In this scheme all the quantizers are assumed to be minimum mean square error (MMSE) quantizers. We note that the quantized values of such quantizers are uncorrelated with the resulting quantization errors.

Scheme.

Encoder. At time t:

- Generates the prediction error

$$\tilde{\boldsymbol{s}}_t \triangleq \boldsymbol{s}_t - \alpha_t \hat{\boldsymbol{s}}_{t-1}\,, \tag{8}$$

where $\hat{\boldsymbol{s}}_{t-1}$, defined in (4), is the previous source reconstruction at the decoder, and $\hat{\boldsymbol{s}}_0 = 0$. A linear recursive relation for $\hat{\boldsymbol{s}}_t$ is provided in the sequel in (9) .[2]

- Generates $\hat{\tilde{\boldsymbol{s}}}_t$, the quantized reconstruction of the prediction error $\tilde{\boldsymbol{s}}_t$, by quantizing $\tilde{\boldsymbol{s}}_t$ using the optimal MMSE quantizer of rate R_t and frame length N.
- Sends $f_t = \hat{\tilde{\boldsymbol{s}}}_t$ over the channel.

Decoder. At time t:

- Receives f_t.
- Recovers a reconstruction $\hat{\tilde{\boldsymbol{s}}}_t$ of the prediction error $\tilde{\boldsymbol{s}}_t$.
- Generates an estimate $\hat{\boldsymbol{s}}_t$ of $\boldsymbol{s}_t$:

$$\hat{\boldsymbol{s}}_t = \alpha_t \hat{\boldsymbol{s}}_{t-1} + \hat{\tilde{\boldsymbol{s}}}_t \,. \tag{9}$$

The optimal achievable distortions $\{D_t\}$ of this scheme for long frame lengths N, are as follows.

Assertion 1 (Inner bound)**.** *Let $\epsilon > 0$, however small. Then, the expected distortion of the scheme at time $t \in [1, T]$ satisfies the recursion*

$$D_t \le \left(\alpha_t^2 D_{t-1} + W_t\right) 2^{-2R_t} + \epsilon, \qquad t = 1, \ldots, T \,, \tag{10a}$$

$$D_0 = 0, \tag{10b}$$

for a large enough N.

Proof: First note that the error between $\boldsymbol{s}_t$ and $\hat{\boldsymbol{s}}_t$, denoted by $\boldsymbol{e}_t$, is equal to

$$\begin{aligned} \boldsymbol{e}_t &\triangleq \boldsymbol{s}_t - \hat{\boldsymbol{s}}_t && \text{(11a)} \\ &= (\tilde{\boldsymbol{s}}_t + \alpha_t \hat{\boldsymbol{s}}_{t-1}) - \left(\alpha_t \hat{\boldsymbol{s}}_{t-1} + \hat{\tilde{\boldsymbol{s}}}_t\right) && \text{(11b)} \\ &= \tilde{\boldsymbol{s}}_t - \hat{\tilde{\boldsymbol{s}}}_t && \text{(11c)} \end{aligned}$$

where (11b) follows from (8) and (9). Thus, the distortion (5) is also the distortion in reconstructing $\tilde{\boldsymbol{s}}_t$.

Using (1), (8) and (11), we express $\tilde{\boldsymbol{s}}_t$ as

$$\begin{aligned} \tilde{\boldsymbol{s}}_t &\triangleq \boldsymbol{s}_t - \alpha_t \hat{\boldsymbol{s}}_{t-1} \\ &= \alpha_t \left(\boldsymbol{s}_{t-1} - \hat{\boldsymbol{s}}_{t-1}\right) + \boldsymbol{w}_t \\ &= \alpha_t \boldsymbol{e}_{t-1} + \boldsymbol{w}_t \,. \end{aligned}$$

Since $\boldsymbol{w}_t$ is independent of $\boldsymbol{e}_{t-1}$, the average power of the entries of $\tilde{\boldsymbol{s}}_t$ is equal to

$$\tilde{S}_t = \alpha_t^2 D_{t-1} + W_t \,.$$

Using the property that the rate–distortion function under mean square error distortion of an i.i.d. source with given variance is upper bounded by that of a white Gaussian source with the same variance (see, e.g., [18, pp. 338–339]), we obtain the following recursion:

$$D_t \le \left(\alpha_t^2 D_{t-1} + W_t\right) 2^{-2R_t},$$

and hence (7) is achievable within an arbitrarily small $\epsilon > 0$, for a sufficiently large N. ∎

[2] $\hat{\boldsymbol{s}}_{t-1} = \mathbb{E}\left[\boldsymbol{s}_{t-1} \middle| f^{t-1}\right]$ and $\alpha_t \hat{\boldsymbol{s}}_{t-1} = \mathbb{E}\left[\boldsymbol{s}_t \middle| f^{t-1}\right]$ are the MMSE estimators of $\boldsymbol{s}_{t-1}$ and $\boldsymbol{s}_t$, respectively, given all the past channel outputs.

B. Impossible (Converse)

We shall now construct an outer bound that coincides with the inner bound of Assert. 1 for large frame lengths N.

Assertion 2 (Outer bound)**.** *Consider the setting of Sec. II. Then, the average achievable distortion D_t at time $t \in [1, T]$ is bounded from below by $D_t \ge D_t^*$, where D_t^* satisfies* (7) *with equality.*

Proof: Let $N \in \mathbb{N}$. We shall prove

$$\begin{aligned} D_t &\ge 2^{-2R_t} \mathbb{E}_{\check{f}^{t-1}} \left[\mathcal{N}\left(\boldsymbol{s}_t \middle| f^{t-1} = \check{f}^{t-1}\right)\right] && \text{(12a)} \\ &\ge D_t^*, \qquad\qquad t = 1, \ldots, T \,, && \text{(12b)} \end{aligned}$$

by induction, where the sequence $\{D_t^*\}$ is defined in (7),

$$\begin{aligned} \mathcal{N}\left(\boldsymbol{s}_t\right) &\triangleq \frac{1}{2\pi \mathrm{e}} 2^{\frac{2}{N} h(\boldsymbol{s}_t)}, \\ \mathcal{N}\left(\boldsymbol{s}_t \middle| f^k = \check{f}^k\right) &\triangleq \frac{1}{2\pi \mathrm{e}} 2^{\frac{2}{N} h\left(\boldsymbol{s}_t \middle| f^k = \check{f}^k\right)} \end{aligned}$$

denote the entropy-power and conditional entropy-power of $\boldsymbol{s}_t$ given $f^k = \check{f}^k$, the expectation $\mathbb{E}_{\check{f}^{t-1}}[\cdot]$ is with respect to $\check{f}^{t-1}$, and the random vector $\check{f}^t$ is distributed the same as f^t.

Basic step **($t = 1$).** First note that, since $\boldsymbol{s}_0 = 0$ and vector $\boldsymbol{w}_1$ consists of i.i.d. Gaussian entries of variance W_1, (12b) is satisfied with equality. To prove (12a), we use the fact that the optimal achievable distortion D_1 for a Gaussian source ($\boldsymbol{s}_1 = \boldsymbol{w}_1$) with i.i.d. entries of power W_1 and rate R_1 is dictated by its rate–distortion function (RDF) [18, Ch. 10.3.2]:

$$D_1 \ge W_1 2^{-2R_1}.$$

Inductive step. Let $k \ge 2$ and suppose (12) is true for $t = k - 1$. We shall now prove that it holds also for $t = k$.

$$\begin{aligned} D_k &= \frac{1}{N} \mathbb{E}\left[\left\|\boldsymbol{s}_k - \hat{\boldsymbol{s}}_k\right\|^2\right] \\ &= \frac{1}{N} \mathbb{E}\left[\mathbb{E}\left[\left\|\boldsymbol{s}_k - \hat{\boldsymbol{s}}_k\right\|^2 \middle| f^{k-1}\right]\right] && \text{(13a)} \\ &= \frac{1}{N} \mathbb{E}_{\check{f}^{k-1}}\left[\mathbb{E}\left[\left\|\boldsymbol{s}_k - \hat{\boldsymbol{s}}_k\right\|^2 \middle| f^{k-1} = \check{f}^{k-1}\right]\right] && \text{(13b)} \\ &\ge \mathbb{E}_{\check{f}^{k-1}}\left[\mathcal{N}\left(\boldsymbol{s}_k \middle| f^{k-1} = \check{f}^{k-1}\right) 2^{-2R_k}\right] && \text{(13c)} \\ &= \mathbb{E}_{\check{f}^{k-1}}\left[\mathcal{N}\left(\alpha_k \boldsymbol{s}_{k-1} + \boldsymbol{w}_k | f^{k-1} = \check{f}^{k-1}\right)\right] 2^{-2R_k} && \text{(13d)} \\ &\ge \Big\{\mathbb{E}_{\check{f}^{k-2}}\left[\mathbb{E}_{\check{f}_{k-1}}\left[\mathcal{N}\left(\alpha_k \boldsymbol{s}_{k-1} | f^{k-1} = \check{f}^{k-1}\right) \middle| \check{f}^{k-2}\right]\right] \\ &\qquad + \mathcal{N}\left(\boldsymbol{w}_k\right)\Big\} 2^{-2R_k} && \text{(13e)} \\ &\ge \Big\{\alpha_k^2 \mathbb{E}_{\check{f}^{k-2}}\left[\mathcal{N}\left(\boldsymbol{s}_{k-1} \middle| f^{k-2} = \check{f}^{k-2},\, f_{k-1}\right)\right] \\ &\qquad + W_k\Big\} 2^{-2R_k} && \text{(13f)} \\ &\ge \Big\{\alpha_k^2 \mathbb{E}_{\check{f}^{k-2}}\left[\mathcal{N}\left(\boldsymbol{s}_{k-1} | f^{k-2} = \check{f}^{k-2}\right)\right] 2^{-2R_{k-1}} \\ &\qquad + W_k\Big\} 2^{-2R_k} && \text{(13g)} \\ &\ge 2^{-2R_k}\left(\alpha_k^2 D_{k-1}^* + W_k\right) && \text{(13h)} \\ &= D_k^*, && \text{(13i)} \end{aligned}$$

where (13a) follows from the law of total expectation, (13b) holds since f^{k-1} and $\check{f}^{k-1}$ have the same distribution, (13c)

follows by bounding from below the inner expectation (conditional distortion) by the rate–distortion function and the Shannon lower bound [18, Ch. 10] — this also proves (12a), (13d) is due to (1), (13e) follows from the entropy-power inequality [18, Ch. 17], (13f) holds since $\boldsymbol{w}_k$ is Gaussian, the scaling property of differential entropies and Jensen's inequality:

$$\begin{aligned}\mathbb{E}_{\check{f}_{k-1}}\Big[2^{\frac{2}{N}h\left(\boldsymbol{s}_{k-1}\left|f^{k-1}=\check{f}^{k-1}\right.\right)}\Big|\check{f}^{k-2}\Big] &\geq 2^{\frac{2}{N}\mathbb{E}_{\check{f}_{k-1}}\left[h\left(\boldsymbol{s}_{k-1}\left|f^{k-1}=\check{f}^{k-1}\right.\right)\right]} \\ &\equiv 2^{\frac{2}{N}h\left(\boldsymbol{s}_{k-1}\left|f^{k-2}=\check{f}^{k-2},\, f_{k-1}\right.\right)},\end{aligned}$$

(13g) follows from the following standard set of inequalities:

$$\begin{aligned}NR_{k-1} &\geq H\left(f_{k-1}\left|f^{k-2}=\check{f}^{k-2}\right.\right) \\ &\geq I\left(\boldsymbol{s}_{k-1}; f_{k-1}\left|f^{k-2}=\check{f}^{k-2}\right.\right) \\ &= h\left(\boldsymbol{s}_{k-1}\left|f^{k-2}=\check{f}^{k-2}\right.\right) \\ &\qquad - h\left(\boldsymbol{s}_{k-1}\left|f^{k-2}=\check{f}^{k-2}, f_{k-1}\right.\right),\end{aligned}$$

(13h) is by the induction hypothesis, and (13i) holds by the definition of $\{D_t^*\}$ as the sequence that satisfies (7) — which also proves (12b). This concludes the proof of (12b) as desired. ∎

Assertion 3 (Outer bound for non-Gaussian noise)**.** *Consider the setting of Sec. II with independent non-Gaussian noise entries $\{w_t\}$. Then, the average achievable distortion D_t at time $t \in [1, T]$ is bounded from below by $D_t \geq D_t^*$, with D_t^* given by the recursion*

$$\begin{aligned}D_t^* &= \left(\alpha^2 D_{t-1}^* + \mathscr{N}\left(w_t\right)\right) 2^{-2R_t}, \\ D_0^* &= 0,\end{aligned}$$

where $\mathscr{N}\left(w_t\right) = \frac{1}{2\pi e} 2^{h(w_t)}$ is the entropy-power of w_t.

Proof: The proof is identical to that of Assert. 2 with W_t replaced by $\mathscr{N}\left(w_t\right)$.[3] ∎

C. Steady State of Asymptotically Stationary Sources

For the asymptotically stationary source in (6), the steady-state average distortion is as follows.

Corollary 1 (Steady state)**.** *Let $\epsilon > 0$, however small. Then, the minimum steady-state distortion is equal to*

$$D_\infty^* = \frac{W 2^{-2R}}{1 - \alpha^2 2^{-2R}} + \epsilon, \tag{15}$$

for a large enough N.

Proof: Note that (15) is a fixed point of (7a) (up to ϵ).

Now since $\alpha < 1$ and $2^{-2R} < 1$, D_t converges to D_∞. This can be easily proved as follows. Assume $D_{t-1} \neq D_\infty$ (otherwise we are already at the fixed point). Then,

$$\begin{aligned}D_t - D_\infty &= \left[\left(\alpha^2 D_{t-1} + W\right) 2^{-2R}\right] - \left[\left(\alpha^2 D_\infty + W\right) 2^{-2R}\right] \\ &= \alpha^2 2^{-2R}\left(D_{t-1} - D_\infty\right),\end{aligned}$$

or equivalently

$$\frac{D_t - D_\infty}{D_{t-1} - D_\infty} = \alpha^2 2^{-2R} < 1.$$

Hence, if $0 \lesseqgtr D_{t-1} - D_\infty$, then

$$0 \lesseqgtr D_t - D_\infty \lesseqgtr D_{t-1} - D_\infty$$

and converges (exponentially fast) to D_∞. ∎

Remark 3. As is evident from the proof, the result of Corol. 1 remains true for any initial value D_0.

IV. Random-Rate Budgets

In this section we generalize the results of Sec. III to random rates $\{r_t\}$ that are independent of each other and of $\{\boldsymbol{w}_t\}$. Rate r_t is revealed to the encoder just before the transmission at time t.

Theorem 2 (Distortion–rate region)**.** *The distortion–rate region of sequential coding with independent rates r^T is given by all distortion tuples D^T that satisfy $D_t \geq D_t^*$ with$D_0^* = 0$ and*

$$D_t^* = \left(\alpha_t^2 D_{t-1}^* + W_t\right) \mathbb{E}\left[2^{-2r_t}\right], \qquad t = 1, \ldots, T\,. \tag{16}$$

Remark 4. An immediate consequence of this theorem and Jensen's inequality is that using packets of a fixed rate that is equal to $\mathbb{E}\left[r_t\right]$ performs better than using random rates.

Proof:

Achievable. Since the achievability scheme in Th. 1 does not use the knowledge of future transmission rates to encode and decode the packet at time t, we have

$$\begin{aligned}d_t &\triangleq \frac{1}{N}\mathbb{E}\left[\left\|\boldsymbol{s}_t - \hat{\boldsymbol{s}}_t\right\|^2 \middle| r^T\right] && \text{(17a)} \\ &= \frac{1}{N}\mathbb{E}\left[\left\|\boldsymbol{s}_t - \hat{\boldsymbol{s}}_t\right\|^2 \middle| r^t\right] && \text{(17b)} \\ &\leq (\alpha_t^2 d_{t-1} + W_t) 2^{-2r_t} + \epsilon, && \text{(17c)}\end{aligned}$$

Taking an expectation of (17c) with respect to r^t and using the independence of r^{t-1} and r_t, we achieve (16).

Impossible. Revealing the rates to the encoder and the decoder prior to the start of transmission can only improve the distortion. Thus, the distortions $\{d_t\}$ conditioned on $\{r_t\}$ (17a) are bounded from below as in Th. 1; by taking the expectation with respect to $\{r_t\}$, we attain the desired result. ∎

For the special case of an asymptotically stationary source (6), the steady-state distortion is given as follows.

Corollary 2 (Steady state)**.** *Assume that the rates $\{r_t\}$ are i.i.d. Let $\epsilon > 0$, however small. Then, the minimum steady-state distortion is equal to*

$$D_\infty = \frac{BW}{1 - \alpha^2 B} + \epsilon \tag{18}$$

for a large enough N, where

$$B \triangleq \mathbb{E}\left[2^{-2r_t}\right].$$

Proof: Note that (18) is a fixed point of (16).

[3] Recall that in the Gaussian setting $\mathscr{N}\left(w_t\right) = \mathrm{Var}\left(\mathrm{w_t}\right) \equiv W_t$.

Since $\alpha < 1$ and $B < 1$, $\mathbb{E}[D_t]$ converges to D_∞. This can be easily proved as follows. Assume $D_{t-1} \neq D_\infty$ (otherwise we are already at the fixed point). Then,

$$\begin{aligned} D_t - D_\infty &= \left[\left(\alpha^2 D_{t-1} + W\right) B\right] - \left[\left(\alpha^2 D_\infty + W\right) B\right] \\ &= \alpha^2 B \left(D_{t-1} - D_\infty\right), \end{aligned}$$

or equivalently

$$\frac{D_t - D_\infty}{D_{t-1} - D_\infty} = \alpha^2 B < 1.$$

Hence, if $0 \leqslant D_{t-1} - D_\infty$, then

$$0 \leqslant D_t - D_\infty \leqslant D_{t-1} - D_\infty$$

and converges exponentially fast to D_∞. ∎

V. Packet Erasures with Instantaneous Feedback

A. One Packet Per Frame

An important special case of the random-rate budget model of Sec. IV is that of packet erasures [13]. Since a packet erasure at time t can be viewed as $r_t = 0$, and assuming that the encoder sends packets of fixed rate R and is cognizant of any packet erasures instantaneously, the packet erasure channel can be cast as the random rate channel of Sec. IV with

$$r_t = b_t R \tag{19a}$$

$$= \begin{cases} R, & b_t = 1 \\ 0, & b_t = 0 \end{cases} \tag{19b}$$

where $\{b_t\}$ are the packet-erasure events, such that $b_t = 1$ corresponds to a successful arrival of the packet f_t at time t, and $b_t = 0$ means it was erased. We further denote by

$$g_t \triangleq b_t f_t \tag{20}$$

the received output where $g_t = 0$ corresponds to an erasure, and otherwise $g_t = f_t$. We assume that $\{b_t\}$ are i.i.d. according to a $\mathscr{B}er(\beta)$ distribution for $\beta \in [0, 1]$.

Remark 5. We shall concentrate on the case of packets of fixed rate R to simplify the subsequent discussion. This way the only randomness in rate comes from the packet-erasure effect. Nevertheless, all the results that follow can be easily extended to random/varying rate allocations to which the effect of packet erasures $\{b_t\}$ is added in the same manner as in (19).

Corollary 3 (Distortion–rate region)**.** *The distortion–rate region of sequential coding with packet erasures and instantaneous feedback is given as in Th. 2 with*

$$\mathbb{E}\left[2^{-2r_t}\right] = 1 - \beta\left(1 - 2^{-2R}\right).$$

Proof: Computing the expectation, we obtain

$$\begin{aligned} \mathbb{E}\left[2^{-2r_t}\right] &= \mathbb{E}\left[2^{-2b_t R}\right] \\ &= \beta 2^{-2R} + (1 - \beta), \end{aligned}$$

as desired. ∎

Corollary 4 (Steady state)**.** *The steady-state distortion is given as in Corol. 2 with*

$$\begin{aligned} B &\triangleq \mathbb{E}\left[2^{-2r_t}\right] \\ &= 1 - \beta\left(1 - 2^{-2R}\right). \end{aligned}$$

B. Multiple Packets Per Frame

In Sec. V-A we assumed one packet (f_t) was sent per each source frame ($\boldsymbol{s}_t$). Instead, one may choose to transmit multiple packets of lower rate per one source frame. Naïve repetition of the same packet trades off diversity for multiplexing in this case [19] and can potentially improve the overall performance.

An improvement over this naïve repetition-based scheme was proposed in [20], where the repetitive transmission of a single compressed description was replaced by multiple descriptions compression [21]–[24].

If we assume the availability of a perfect instantaneous feedback after each packet, a further improvement can be achieved by noting that this scenario falls again in the random-rate budget framework of Sec. IV.

Specifically, if we assume the use of K packets of equal rate R/K (and hence a total rate of R), the rate probability distribution amounts to

$$r_t = \frac{b_t}{K} R$$

with b_t denoting the number of successful packet arrivals at time t, corresponding to source frame $\boldsymbol{s}_t$. Assuming that the erasure events of all packets are i.i.d. with probability $1 - \beta$ implies that $\{b_t\}$ are i.i.d. according to a Binomial distribution $\mathscr{B}(K, \beta)$.

Interestingly, the optimal number of packets depends on the (total) rate R and packet successful arrival probability β, and is determined by the number that minimizes $\mathbb{E}\left[2^{-r_t}\right]$. This is demonstrated in Fig. 2.

Remark 6. We only considered uniform rate allocations for all the packets. Clearly, one can generalize the same approach to non-uniform packet rates.

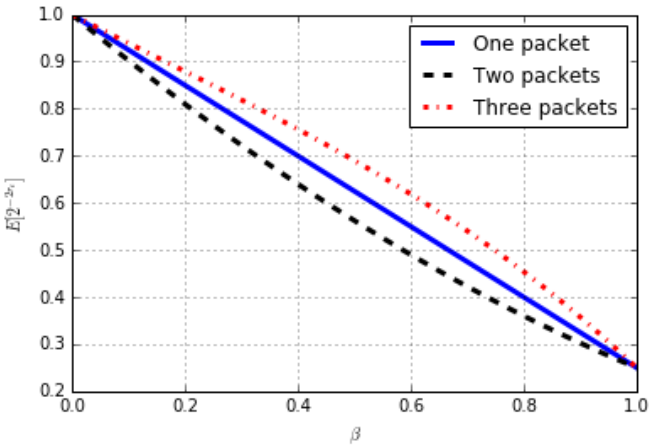
Remark 7. In practice one might expect longer packets to be prone to higher erasure probability. This can be taken into account when deciding on the K that minimizes $\mathbb{E}\left[2^{-2r_t}\right]$.

VI. Packet Erasures with Delayed Feedback

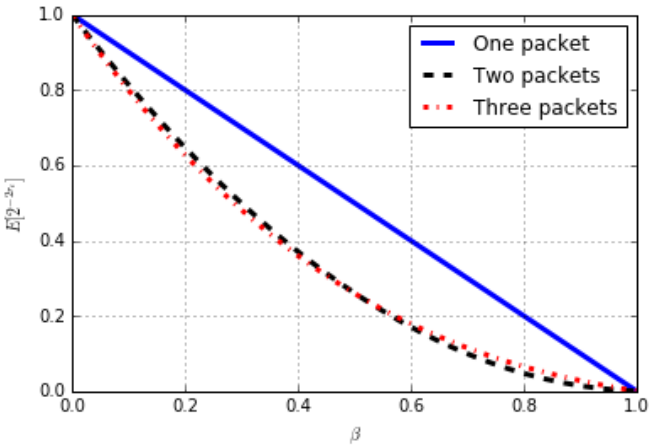
In this section we consider the case of i.i.d. packet erasures with a delayed-by-one output feedback, i.e., the case where at time t, the encoder does know whether the last packet arrived or not (does not know b_{t-1}), but knows the erasure pattern of all preceding packets (knows b^{t-2}). The encoder (3) and decoder (4) mappings can be written as [recall the definition of $g_t \triangleq b_t f_t$ in (20)]:

$$\begin{aligned} f_t &= \mathscr{F}_t\left(\boldsymbol{s}^t, g^{t-2}\right), \\ \hat{\boldsymbol{s}}_t &= \mathscr{G}_t\left(g^t\right). \end{aligned}$$

To that end, we recall the following result by Perron *et al.* [15, Th. 2], which is a specialization to the jointly Gaussian



(a) $R = 1$



(b) $R = 5.5$

Fig. 2: Evaluation of 2^{-r_t} for $K = 1, 2$ and 3 packets, all possible values of $\beta \in [0, 1]$, and two (total) rates $R = 1$ and 5.5.

case of the result by Kaspi [14, Th. 1], who established the rate–distortion region of lossy compression with two-sided SI where the SI may or may not be available at the decoder.[4]

Remark 8. Kaspi's result [14, Th. 1] can also be viewed as a special case of [16] with some adjustments; see [17].

Theorem 3 ([15])**.** *Let $\boldsymbol{s}$ be an i.i.d. zero-mean Gaussian source of power S, which is jointly Gaussian with SI $\boldsymbol{y}$, which is available at the encoder and satisfies $\boldsymbol{s} = \boldsymbol{y} + \boldsymbol{z}$ where $\boldsymbol{z}$ is an i.i.d. Gaussian noise of power Z that is independent of $\boldsymbol{y}$. Denote by $\hat{\boldsymbol{s}}^+$ and $\hat{\boldsymbol{s}}^-$ the reconstructions of $\boldsymbol{s}$ with and without the SI $\boldsymbol{y}$, and by D^+ and D^- their mean squared error distortion requirements, respectively. Then, the smallest rate required to achieve these distortions is given by*

$$R^{\text{Kaspi}}(S, Z, D^-, D^+)$$
$$= \begin{cases} 0, & D^- \geq S \text{ and } D^+ \geq Z \\ \frac{1}{2}\log\left(\frac{S}{D^-}\right), & D^- < S \text{ and } D^+ \| S \geq D^- \| Z \\ \frac{1}{2}\log\left(\frac{Z}{D^+}\right), & D^+ < Z \text{ and } D^- \geq D^+ + S - Z \\ \frac{1}{2}\log\left(\frac{S}{D^- - \Delta^2}\right), & \begin{cases} D^- < S \text{ and } D^+ \| S < D^- \| Z \\ \text{and } D^- < D^+ + S - Z \end{cases} \end{cases}$$

where $a \| b \triangleq \frac{ab}{a+b}$ denotes the harmonic mean of a and b, and

$$\Delta \triangleq \frac{\sqrt{(S-Z)(S-D^-)}D^+ - \sqrt{(Z-D^+)(D^- - D^+)}S}{\sqrt{Z}\,(S - D^+)}\,.$$

Remark 9. Surprisingly, as observed by Perron *et al.* [15], if the side-information signal $\boldsymbol{y}$ is not available at the encoder — corresponding to the case considered in [16] and [14, Th. 2] — the required rate can be strictly higher than that in Th. 3. This is in stark contrast to the case where the side-information is never available at the encoder and the case where the side-information is always available at the decoder studied by Wyner and Ziv [25], [26]. Knowing the SI at the encoder allows to (anti-)correlate the noise $\boldsymbol{z}$ with the quantization error — some thing that is not possible when the SI is not available at the encoder, as the two noises must be independent in that case. This allows for some improvement, though a modest one, as implied by the results for the dual channel problem [27, Prop. 1], [28].

In our case, at time t, the previous packet f_{t-1} will serve as the SI. Note that it is always available to the encoder; the decoder may or may not have access to it, depending whether the previous packet arrived or not. Since the feedback is delayed, during the transmission of the current packet f_t the encoder does not know whether the previous packet was lost.

The tradeoff between D^+ and D^- for a given rate R will be determined by the probability of a successful packet arrival β.

Scheme (Kaspi-based)**.**

Encoder. At time t:

- Generates the prediction error

$$\tilde{\boldsymbol{s}}_t \triangleq \boldsymbol{s}_t - \alpha_t^2 \hat{\boldsymbol{s}}_{t-2}\,.$$

- Generates f_t by quantizing the prediction error $\tilde{\boldsymbol{s}}_t$ as in Th. 3, where f_{t-1} is available as SI at the encoder and possibly at the decoder (depending on b_{t-1}) using the optimal quantizer of rate R and frame length N that minimizes the averaged over b_{t-1} distortion:

$$D_t^{\text{Weighted}} = \beta D_t^+ + (1-\beta) D_t^-\,; \tag{21}$$

more precisely, since the encoder does not know (b_{t-1}, b_t) at time t:

  - Denote the reconstruction of $\tilde{\boldsymbol{s}}_t$ at the decoder from f_t and g^{t-1} — namely given that $b_t = 1$ — by $Q_t(\tilde{\boldsymbol{s}}_t)$, and the corresponding distortion by D_t^{Weighted}.

[4] We use a backward channel to represent the SI $\boldsymbol{s} = \boldsymbol{y} + \boldsymbol{z}$, as opposed to the forward channel $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{z}$ used in [15], [16].

– Denote the reconstruction from f_t and g^{t-2} — namely given that $b_t = 1$ and $b_{t-1} = 0$ — by $Q_t^-(\tilde{\boldsymbol{s}}_t)$, and the corresponding distortion by D_t^-.
– Denote the reconstruction from (f_{t-1}, f_t) and g^{t-2} — namely given that $b_t = 1$ and $b_{t-1} = 1$ — by $Q_t^+(\tilde{\boldsymbol{s}}_t)$, and the corresponding distortion by D_t^+.

Then, the encoder sees $\alpha_t Q_{t-1}^+(\tilde{\boldsymbol{s}}_{t-1})$ as possible SI available at the decoder to minimize D_t^{Weighted} as in (21).

- Sends f_t over the channel.

Decoder. At time t:

- Receives g_t.
- Generates a reconstruction $\hat{\tilde{\boldsymbol{s}}}_t$ of the prediction error $\tilde{\boldsymbol{s}}_t$:

$$\hat{\tilde{\boldsymbol{s}}}_t = \begin{cases} Q_t^+(\tilde{\boldsymbol{s}}_t), & b_t = 1, b_{t-1} = 1 \\ Q_t^-(\tilde{\boldsymbol{s}}_t), & b_t = 1, b_{t-1} = 0 \\ 0, & b_t = 0 \end{cases} \tag{22}$$

- Generates an estimate $\hat{\boldsymbol{s}}_t$ of $\boldsymbol{s}_t$:

$$\hat{\boldsymbol{s}}_t = \alpha_t \hat{\boldsymbol{s}}_{t-1} + \hat{\tilde{\boldsymbol{s}}}_t \,.$$

This scheme is the optimal greedy scheme whose performance is stated next, in the limit of large N.

Theorem 4. *Let $\epsilon > 0$, however small. Then, for a large enough N, the expected distortion of the scheme at time $t \in [2, T]$ given $(b_1, \ldots, b_t)$ satisfies the recursion*

$$D_t = \begin{cases} D_t^+ + \epsilon, & b_t = 1, b_{t-1} = 1 \\ D_t^- + \epsilon, & b_t = 1, b_{t-1} = 0 \\ \alpha_t^2 D_{t-1} + W + \epsilon, & b_t = 0 \end{cases}$$
$$D_1 = D_1^+ = D_1^- = W_t 2^{-b_1 2R} + \epsilon,$$

where D_t^+ and D_t^- are the distortions that minimize

$$D_t^{\text{Weighted}} = \beta D_t^+ + (1-\beta) D_t^- \,,$$

such that the rate of Th. 3 satisfies

$$R^{\text{Kaspi}}(\alpha_t D_{t-1}^- + W, \alpha_t D_{t-1}^+ + W, D_t^-, D_t^+) = R.$$

The proof is again the same as that of Ths. 1 and 2, with $\hat{\tilde{\boldsymbol{s}}}_t$ generated as in (22).

Remark 10. Here, in contrast to the case of instantaneous feedback, evaluating the average distortions $\{D_t\}$ in explicit form (recall Corol. 3) is much more challenging. We do it numerically, instead.

Somewhat surprisingly, the loss in performance of the Kaspi-based scheme due to the feedback delay is rather small compared to the scenario in Sec. V where the feedback is available instantaneously, for all values of β.[5] This is demonstrated in Fig. 3, where the perfomances of these schemes are compared along with the performances of the following three simple schemes for $\alpha_t \equiv 0.7, W \equiv 1, \beta = 0.5, R = 2$

[5]For β values close to 0 or 1, the loss becomes even smaller as in these cases using the scheme of Sec. V that assumes that the previous packet arrived or was erased, respectively, becomes optimal.

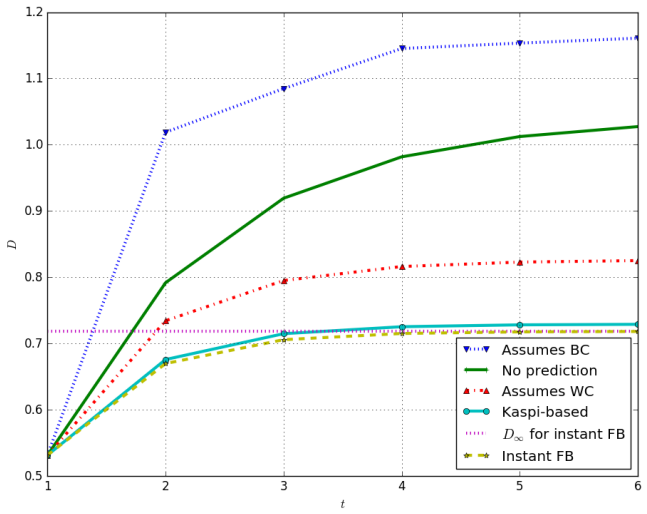


Fig. 3: Distortions D_t as a function of the time t of the various schemes presented in this section, along with that of the instantaneous-feedback scheme of Sec. V, for $\alpha = 0.7$, $W = 1$, $\beta = 0.5$ and $R = 2$.

(we derive their performance for the special case of an asymptotically stationary source):

- **No prediction:** A scheme that does not use prediction at all, as if the source samples were independent. This scheme achieves a distortion of

$$D_t = \beta S_t 2^{-2R} + (1-\beta) S_t, \qquad t = 1, \ldots, T \,,$$

where S_t is the power of the entries of $\boldsymbol{s}_t$ as given in (2).

- **Assumes worst case (WC):** Since at time t the encoder does not know b_{t-1}, a "safe" way would be to work as if $b_{t-1} = 0$. This achieves a distortion of

$$D_t = \left[\alpha^4 D_{t-2} + (1+\alpha^2) W\right] \left[\beta 2^{-2R} + (1-\beta)^2\right] + \beta(1-\beta)(\alpha^2 D_{t-1} + W), \qquad t = 2, \ldots, T \,,$$
$$D_0 = 0, \qquad D_1 = W 2^{-2R}.$$

- **Assumes best case (BC):** The optimistic counterpart of the previous scheme is that which always works as if $b_{t-1} = 1$. This scheme achieves a distortion of

$$D_t = \beta \left[\alpha^2 D_{t-1|t-2} 2^{-2R} + W\right] \left[\beta 2^{-2R} + (1-\beta)\right] + (1-\beta)\left[\alpha^2 D_{t-1|t-2} + W\right], \quad t = 2, \ldots, T \,,$$
$$D_{t-1|t-2} \triangleq \alpha^2 D_{t-2} + W, \qquad t = 2, \ldots, T \,,$$
$$D_0 = 0, \quad D_1 = W 2^{-2R}.$$

VII. Discussion: Feedback with Larger Delays

To extend the scheme of Sec. VI for larger delays, a generalization of Th. 3 is needed. Unfortunately, the optimal rate–distortion region for more than two decoders remains an open problem and is only known for the case when the source and the possible SIs form a Markov chain ("degraded"). Nonetheless, achievable regions for multiple decoders have been proposed in [16], which can be used for the construction of schemes that accommodate larger delays.

ACKNOWLEDGMENT

The authors thank Yu Su from Caltech for valuable discussions.

REFERENCES

[1] H. Viswanathan and T. Berger, "Squential coding of correlated sources," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 236–246, Jan. 2000.

[2] N. Ma and P. Ishwar, "On delayed sequential coding of correlated sources," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3763–3782, 2011.

[3] ——, "Erratum to "on delayed sequential coding of correlated sources"," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, p. 4074, June 2012.

[4] E.-H. Yang, L. Zheng, and D.-K. He, "Rate distortion theory for causal video coding: Characterization, computation algorithm, and comparison," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5258–5280, 2011.

[5] ——, "On the information theoretic performance comparison of causal video coding and predictive video coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1428–1446, Mar. 2014.

[6] M. Torbatian and E.-H. Yang, "Causal coding of multiple jointly Gaussian sources," in *Proc. Annual Allerton Conf. on Comm., Control, and Comput.*, Monticello, IL, USA, Oct. 2012, pp. 2060–2067.

[7] M. Eslamifar, "On causal video coding with possible loss of the first encoded frame," Master's thesis, University of Waterloo, Waterloo, Ontario, Canada, 2013.

[8] L. Song, J. Chen, J. Wang, and T. Liu, "Gaussian robust sequential and predictive coding," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3635–3652, June 2013.

[9] H. C. Huang, W. H. Peng, and T. Chiang, "Advances in the scalable amendment of H.264/AVC," *IEEE Comm. Magazine*, vol. 45, no. 1, pp. 68–76, Jan. 2007.

[10] Y.-Z. Huang, Y. Kochman, and G. W. Wornell, "Causal transmission of colored source frames over packet erasure channel," in *Proc. Data Comp. Conf. (DCC)*, Snowbird, UT, USA, Mar. 2010, pp. 129–138.

[11] F. Etezadi, A. Khisti, and M. Trott, "Zero-delay sequential transmission of Markov sources over burst erasure channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4584–4613, Aug. 2014.

[12] F. Etezadi, A. Khisti, and J. Chen, "A truncated prediction framework for streaming over erasure channels," *IEEE Trans. Inf. Theory*, Submitted Jul. 2014, Revised Oct. 2016.

[13] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Trans. Auto. Control*, vol. 54, no. 2, pp. 243–255, Feb. 2009.

[14] A. H. Kaspi, "Rate–distortion when side-information may be present at the decoder," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2031–2034, Nov. 1994.

[15] E. Perron, S. Diggavi, and I. E. Telatar, "On the role of encoder side-information in source coding for multiple decoders," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Seattle, WA, USA, July 2006, pp. 331–335.

[16] C. Heegard and T. Berger, "Rate–distortion when side information may be absent," *IEEE Trans. Inf. Theory*, vol. 31, pp. 727–734, Nov. 1985.

[17] A. Khina and U. Erez, "Source coding with composite side information at the decoder," in *Proc. IEEE Conv. Electrical and Electron. Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory, Second Edition*. New York: Wiley, 2006.

[19] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. U.K: Cambridge Univ. Press, 2005.

[20] J. Ostergaard and D. Quevedo, "Multiple descriptions for packetized predictive control," *EURASIP J. Advances in Sig. Proc.*, vol. 2016, no. 45, Apr. 2016.

[21] H. Witsenhausen, "On source networks with minimal breakdown degradation," *Bell Sys. Tech. Jour.*, vol. 59, pp. 1083–1087, July-Aug. 1980.

[22] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple descriptions," *Bell Sys. Tech. Jour.*, vol. 59, pp. 1417–1426, Oct. 1980.

[23] L. H. Ozarow, "On the source coding problem with two channels and three receivers," *Bell Sys. Tech. Jour.*, vol. 59, pp. 1909–1922, 1980.

[24] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inf. Theory*, vol. 28, no. 6, pp. 851–857, Nov. 1982.

[25] A. D. Wyner and J. Ziv, "The rate–distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[26] A. D. Wyner, "The rate–distortion function for source coding with side information at the decoder — II: General sources," *Information and Control*, vol. 38, pp. 60–80, 1978.

[27] R. Zamir and U. Erez, "A Gaussian input is not too bad," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1362–1367, Jun. 2004.

[28] T. Philosof and R. Zamir, "The cost of uncorrelation and noncooperation in MIMO channels," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 3904–3920, Nov. 2007.